Supplemental Information

NEXT-GENERATION SEQUENCING AND VARIANT ANALYSIS

Solution hybridization exome capture was performed by using the SureSelect Human All Exon System (Agilent Technologies, Santa Clara, CA), which uses biotinylated RNA baits to hybridize to sequences that correspond to exons.[1] Manufacturer's protocol version 1.0 compatible with Illumina (Illumina Inc, San Diego, CA) paired-end sequencing was used, with the exception that DNA fragment size and quality was measured by using a 2% agarose gel stained with Sybr Gold rather than using an Agilent Bioanalyzer. The manufacturer's specifications state that the capture regions total ~38 Mb. This kit covers 1.22% of the human genome, corresponding to the Consensus Conserved Domain Sequences database and >1000 noncoding RNAs. Flowcell preparation and end read sequencing were carried out per protocol for the GAIIx sequencer (Illumina Inc).[2] Two 76-bp paired-end lanes on a GAIIx flowcell were used per exome sample to generate sufficient reads to generate the aligned sequence. Image analysis and base calling on all lanes of data were performed by using Illumina Genome Analyzer Pipeline software (GAPipeline versions 1.4.0 or greater; Available at: www.illumina.com/software/genome_analyzer_software.ilmn) with default parameters.

READ MAPPING, VARIANT CALLING, AND ANNOTATION

Reads were aligned to a human reference sequence (University of California Santa Cruz [UCSC] assembly hg18, National Center for Biotechnology Information build 36) by using "efficient large-scale alignment of nucleotide databases." Reads that aligned uniquely were grouped into genomic sequence intervals of ~100 kilobases (kb); reads that failed to align were binned with their paired-end mates. Reads in each bin were subjected to a Smith-Waterman–based local alignment algorithm, *cross_match* by using the parameters –minscore 21 and –masklevel 0 to their respective 100-kb genomic sequence (http://www.phrap.org). A total of 6 gigabases of high-confidence mappable sequence data were generated in autosomal targeted regions per individual. Genotypes were called at all positions with high-quality sequence bases (Phred-like Q20 or greater) by using a Bayesian algorithm (MPG).[3] Genotypes with an MPG score ≥10 demonstrate >99.89% concordance with SNP Chip data. The targeted regions included the exons of 17 134 genes, with a total of 36 025 890 bases in the human genome. We were able to capture and sequence 93.1% of the exome (as defined by UCSC known gene annotations) in twin A and 92.4% of the exome in twin B. Annotation of coding single nucleotide variants (SNVs) was based on UCSC's "known genes" dataset. Missense variants were sorted by the degree of severity of functional disruption prediction by using Conserved Domain-based Prediction.[4]

By using Conserved Domain-based Prediction, sequence variants relative to human reference sequence (hg18), namely SNVs and short deletion-insertion variants (DIVs), were first identified by using MPG. SNVs and DIVs were classified by a custom suite of annotation scripts (Protein Integrated ANNOtation [PIANNO]) as those in intronic, ultratranslated regions, or within coding regions. The software further computed the consequence of SNVs as either missense, nonsense, or silent, splice-site affecting, or coding frame altering for DIVs, respectively. The functional consequence of the missense variants was scored based on the degree of conservation at the substitution site (δ-score is a measure of deviation from the reference amino-acid). The more negative the δ-score, the more deleterious the prediction. Finally, variants detected in dbSNP (version 130; Available at: www.ncbi.nlm.nih.gov/projects/SNP/) were not analyzed further in this context.[4]

SUPPLEMENTAL REFERENCES

1. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*. 2009;27(2):182–189
2. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53–59
3. Teer JK, Bonnycastle LL, Chines PS, et al; NISC Comparative Sequencing Program. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res*. 2010;20(10):1420–1431
4. Johnston JJ, Teer JK, Cherukuri PF, et al; NIH Intramural Sequencing Center (NISC). Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet*. 2010;86(5):743–748